Published in Elsevier *Expert Systems with Applications* in 2024. This is the accepted version from https://www.beierle.de.

Predicting adherence to ecological momentary assessments

Felix Beierle[a,b,*], Wepan Chada[c], Akiko Aizawa[a] and Rüdiger Pryss[b]

[a]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
[b]Institute of Clinical Epidemiology and Biometry (ICE-B), University of Würzburg, Sanderring 2, Würzburg, 97070, Germany
[c]Service-centric Networking, Technische Universität Berlin, Ernst-Reuter-Platz 7, Berlin, 10587, Germany

ARTICLE INFO

Keywords:
adherence
ecological momentary assessment
mobile sensing
machine learning
prediction

ABSTRACT

Smartphones allow for prompting users with a short questionnaire about their current subjective experience, a technique often called Ecological Momentary Assessments. One of the biggest challenges for such studies is a lack of adherence, diminishing the benefits for both user and researcher. Being able to predict if a user is going to stop answering the questionnaire prompts would be beneficial for researchers and developers. This would allow for, for example, specifically addressing those users, or for over-sampling populations at higher risk of dropping out of a study. In this work, based on an observational study of the general population, we analyzed data from almost 1,000 users. The data include a large variety of sensor data from the users' smartphones. We utilized machine learning to predict adherence on a day-to-day level, as well as predict adherence based on participant data after on-boarding. For day-to-day prediction, the best performing model was a model based on metadata features (days since first questionnaire was filled out, days since the last questionnaire was filled out, number of filled-out questionnaires, days since app installation), yielding an area under the precision-recall curve of 0.89. The inclusion of sensor data did not improve the model's performance, indicating that the high cost of collecting and processing sensor data is not worth the benefits for predicting fill-out behavior. Predicting at sign-up if a user will adhere to a questionnaire prompt at least once was better than chance, but further studies are needed.

1. Introduction

Smartphones can be used to query app users in situ and prompt them to fill out a short questionnaire about their current experience. This idea is often referred to as ecological momentary assessment (EMA) (Stone and Shiffman, 1994; Shiffman, Stone and Hufford, 2008). The questionnaires' content can relate to mood or symptom tracking, or be used for collecting data for researching changing personality states (Schlee, Pryss, Probst, Schobel, Bachmeier, Reichert and Langguth, 2016; Beierle, Tran, Allemand, Neff, Schlee, Probst, Pryss and Zimmermann, 2018b). *Mobile sensing* refers to using the sensors of a smartphone and recording and storing measurements (Burke, Estrin, Hansen, Parker, Ramanathan, Reddy and Srivastava, 2006; Beierle, Tran, Allemand, Neff, Schlee, Probst, Pryss and Zimmermann, 2018a; Boubiche, Imran, Maqsood and Shoaib, 2019; Virginia Anikwe, Friday Nweke, Chukwu Ikegwu, Adolphus Egwuonwu, Uchenna Onu, Rita Alo and Wah Teh, 2022). Combining mobile sensing with EMA allows for the combination of objective sensor data with subjective experiences (Beierle, Tran, Allemand, Neff, Schlee, Probst, Zimmermann and Pryss, 2020; Kraft, Schlee, Stach, Reichert, Langguth, Baumeister, Probst, Hannemann and Pryss, 2020; Beierle, 2021; Beierle, Matz and Allemand, 2023), enabling deeper research in domains such as medicine or psychology. One of the core issues of EMA systems is that many users only fill out just a few or only one questionnaire and then stop using the app. This diminishes the benefits for the user himself/herself, as well as the researcher. Sticking to filling out a questionnaire regularly is often referred to as *adherence*.

However, it is difficult for researchers and developers of EMA apps to predict (non-)adherence, i.e., to predict which users are going to drop out and stop responding to prompts. While there has been some related work in this field, to the best of our knowledge, we are the first to report results from a study that (1) specifically tries to predict fill-out behavior on a day-to-day basis, as well as (2) utilizes mobile sensing data for machine-learning-based predictions. There are several benefits for predictions about adherence. For future studies, researchers can gauge what users to expect and potentially over-sample those that are predicted to drop out. Given that developing and maintaining an app is costly and time-consuming, we believe that answering questions about predicting adherence is significant and relevant. In more general terms, adherence prediction can help researchers and developers better understand their user base to potentially specifically address users at risk of stopping the use of an app.

In this article, we present our study based on a longitudinal observational study of the general population. We collected the data with our app TYDR – Track Your Daily Routine (Beierle et al., 2018b), which was free to use and prompted a daily questionnaire about personality states, containing questionnaires about the user's daily experiences and behaviors. The purpose of this study, and thus our main contribution, is to answer the following research questions:

RQ1 To what extend can we predict if a user is going to fill out a daily questionnaire in the evening?

*Corresponding author
felix.beierle@uni-wuerzburg.de (F. Beierle); wepan.chada@gmail.com (W. Chada); aizawa@nii.ac.jp (A. Aizawa); ruediger.pryss@uni-wuerzburg.de (R. Pryss)
ORCID(s): 0000-0003-2702-9893 (F. Beierle); 0000-0003-1522-785X (R. Pryss)

RQ2 To what extend can we predict if users are going to fill out at least one daily questionnaire, based on their information collected during on-boarding?

The rest of the paper is structured as follows. In Section 2, we give an overview about related work. In Section 3, we detail the methods used and describe in detail the data used in our study. In Sections 4 and 5, we present the results and discuss our findings.

2. Related Work

In the related work, there are studies that cover different groups of EMA app users. This includes specific patient groups as well as the general population. The length of the studies also differs, from one week to voluntary app usage over several years. Common topics covered in related work are (a) what app user/study participant characteristics correlate with adherence, (b) what factors recorded during the study predict adherence to the rest of the study, and (c) the influence of prompting frequency on adherence.

In Table 1, we give an overview about the population, timeframe, and data used in related studies. Most studies are done with specific patient groups instead of the general population. Most studies have between 68 to 260 participants. There are two outliers with 1184 and 1292 participants, neither studied the general population. The timeframes of most studies is between 7 and 30, with outliers where the app usage was voluntary and no end was defined, making it possible for users to use the app for several years. None of the existing studies used sensor data. Only one of the studies uses machine learning approaches (Schleicher, Unnikrishnan, Neff, Simoes, Probst, Pryss, Schlee and Spiliopoulou, 2020); all others use statistical analyses.

Wen, Schneider, Stone and Spruijt-Metz (2017) surveyed the literature on mobile EMA compliance in studies with children and adolescents. They found that app users in clinical settings tend to adhere better with higher prompting frequency (6 and more), while non-clinical studies reported higher adherence for a lower prompting frequency (2-3 times a day).

Colombo, Cipresso, Alvarez, Palacios, Riva and Botella (2018) analyzed several studies with patients suffering from major depressive disorder. They report that the mean adherence was higher when prompting the participant less than 8 times a day.

Gershon, Kaufmann, Torous, Depp and Ketter (2019) analyzed patient characteristics that impact EMA adherence in youth with bipolar disorder. They found that adherence was worse in bipolar youth compared to the healthy control group, with "lifetime suicide attempts and higher current mood elevation" being factors for predicting worse adherence.

Williams-Kerver, Schaefer, Hazzard, Cao, Engel, Peterson, Wonderlich and Crosby (2021) investigated adherence to EMA prompts in adults with binge-eating disorder. They report that person-level characteristics did not predict adherence, while affect, hunger, and signals later in the day predicted adherence.

Murray, Yang, Zhu, Speyer, Brown, Eisner and Ribeaud (2023a) analyzed respondent characteristics associated with adherence in the general population. Their results suggest that traits related to a lack of self-regulation and anti-social behavior/emotions may predict lower EMA adherence. In another paper, again on a sample of the general population, they report that stress and negative affect predicted non-response (Murray, Brown, Zhu, Speyer, Yang, Xiao, Ribeaud and Eisner, 2023b).

Klaus, Peek, Quynh, Sutherland, Selvam, Moore, Depp and Eyler (2022) conducted a study on mobile survey engagement by older adults and report that "[a]dherence was predicted by education status, study participation duration, reaching the study midpoint and time between study start and enrollment."

Jones, Moore, Pinkham, Depp, Granholm and Harvey (2021a) report on both patients and healthy control groups that early adherence predicted study-long adherence. Adherence did not correlate with mood, study length, nor prompting frequency.

Jones, Hue, Kelly, Barnett, Henderson and Sengupta (2021b) analyzed data from a study with patients suffering from a rheumatological condition. Filling out EMA prompts in the evening was associated with higher adherence, as was older age.

Schleicher et al. (2020) analyzed the adherence of app users that report suffering from tinnitus. Voluntary app users could fill out a questionnaire multiple times a day. Data were collected over a longer period of time (almost three years between 2014-2017). The app did not collect sensor data and adherence was recorded based on interaction with the questionnaire, i.e., filling out at least one question of the questionnaire, and fill-out continuity, i.e, how many days in a row the questionnaire was filled out. Schleicher et al. found that fill-out behavior from the first days of usage can help predict continued adherence. At least on their dataset, they could not predict adherence based on data at registration.

Overall, the findings from each cited paper seem to not necessarily generalize well to other study samples. For higher adherence, some works seem to suggest higher prompting frequencies, some seem to suggest lower frequencies. Some studies report correlations with user characteristics like age or prompting time, while others report no correlations with user characteristics. Some studies report that data from during the study (current adherence, mood, affect) can help in the prediction for the rest of the study. Our work differs in some key aspects: none of the cited papers built a machine learning model for the day-to-day prediction of adherence. Additionally, to the best of our knowledge, we are the first to utilize sensor data for trying to predict adherence to EMA prompts.

Table 1
Related work overview about survey and studies focusing on adherence to EMAs.

Reference	Population	Timeframe	Survey data	Sensor data
Wen et al. (2017) survey: 42 studies	children and adolescents; average n = 98.81	average 11.4 days for non-clinical; 16.3 days for clinical studies	y	-
Colombo et al. (2018) survey: 13 studies	people with major depressive disorder; average n = 68.38	average 10.53 days	y	-
Gershon et al. (2019)	youth with bipolar disorder (n = 39), healthy control (n = 47)	21 days	y	-
Williams-Kerver et al. (2021)	adults with binge-eating disorder; n = 110	7 days	y	-
Murray et al. (2023a)	general population (young adults); n = 255	14 days	y	-
Murray et al. (2023b)	general population (young adults); n = 260	14 days	y	-
Klaus et al. (2022)	general population (older adults); n = 95	4 assessments periods of 16 days each	y	-
Jones et al. (2021a)	study 1: schizophrenia (n = 106), healthy control (n = 76) study 2: schizophrenia (n = 104), bipolar illness (n = 76)	study 1: 7 days study 2: 30 days	y	-
Jones et al. (2021b)	people with a specific rheumatological condition; n = 1184	up to 593 days	y	-
Schleicher et al. (2020)	people with tinnitus; n = 1292	between April 2014 and February 2017	y	-
this work	general population; study 1: n = 942 study 2: n = 794	between October 2018 and May 2021	y	study 1: y study 2: -

3. Methods

In this section, we report about the data collection, data cleaning and processing, the final datasets, and the machine learning approaches and evaluation metrics.

3.1. Data Collection

We collected the data with our Android app TYDR – Track Your Daily Routine (Beierle et al., 2018b). TYDR collected a large variety of sensor data via mobile sensing and poses questionnaires related to demographics and personality. This includes rather static *personality traits*, captured by the Big Five model (McCrae and John, 1992). These traits fluctuate within persons across time (Fleeson, 2001). To capture these fluctuations, sometimes referred to as *personality states*, we employed the Personality Dynamics Diary (PDD) (Zimmermann, Woods, Ritter, Happel, Masuhr, Jaeger, Spitzer and Wright, 2019), asking the user questions about his/her experiences of daily situations and behaviors. The PDD could be filled out every evening between 6pm and 2am.

The ethics commission of the Technical University of Berlin approved the study on May 23, 2018 with code BEI_01_20180115. We released TYDR in Google Play in October 2018. TYDR is publicly available on Google Play. Between October 2018 and June 2019, the app showed a study asking participants to fill out PDD for 21 days. Regardless of joining the study, the users could fill out PDD as many times as they chose to. For the study reported in this article, we collected data from October 2018 to May 2021.

3.2. Data Cleaning and Processing

We report the specifics of preparing the datasets for conducting our experiments.

3.2.1. Daily prediction of filling out PDD

We recorded app usage statistics via the *UsageEvents*[1] available in Android, that developers can access as long as the user gives the associated special permission (cf. Beierle et al., 2018a, 2020). Each event is recorded with a timestamp and a specific *Event*[2] indicating, for example, an app coming into the foreground. Based on the UsageEvents, we then calculated how long each app was used, i.e., was in the foreground. We sorted all apps into 74 categories. *Social Networks*, *Messaging*, and *TV/Video-Apps* were the overall app categories that had the longest usage sums over all users.

[1] https://developer.android.com/reference/android/app/usage/UsageEvents, accessed 2023-05-23.

[2] https://developer.android.com/reference/android/app/usage/UsageEvents.Event, accessed 2023-05-23.

We calculated the app usage in seconds for each category for each user. For data cleaning, we removed outliers that exist likely due to faulty recordings. To avoid target leakage, we set up the use of TYDR as its own category and removed its usage for 6pm to 2am, i.e., the time during which the PDD could be filled out. Overall, we had less than 2% missing values. The matrix consisting of app usage categories as columns and each available day/user combination as rows is very sparse. We imputed the missing values with 0.

In addition to the app usage data, we collected several sources of sensor data (cf. Beierle et al., 2018a). For the present study, we used metadata about phone calls, WiFi connection status, battery level and charger connection status, display state (on/off), photo metadata (how many pictures taken including the orientation of the phone), and steps count and duration. Because of the fragmentation of Android devices with many different device manufacturers and OS adaptation by the manufacturers, not all devices yield the same amount of data from each sensor with the same reliability. That is why we observed many missing values. We dropped those columns where we had less than 50% of non-missing values. Imputation of missing values typically works by inferring values based on non-missing values. Thus, we drop those features where we do not have enough existing values to reasonably infer. On the other hand, we want to avoid dropping features that, despite having missing values, are beneficial for the prediction. While we had to drop some features, we still have detailed information about the user's day and his/her activities, by having several features containing information about WiFi, battery, display state, call metadata, photo metadata, step count and duration, and app usage.

In addition to the sensor data features described above, we can derive additional features for the daily prediction, based on the past behavior of the user. We call these features *metadata features*. We defined four metadata features:

MD1 Days since first PDD – The number of days that have passed since the first PDD was filled out by the user.

MD2 Days since last PDD – The number of days that have passed since the last PDD was filled out by the user.

MD3 Number of PDDs to date – The total number of times the user filled out PDD until the current date.

MD4 Days since app installation – The number of days since the user started using TYDR.

MD1-4 are discrete variables where most values are 0. For MD1, MD2, and MD3, the value 0 indicates that for the given user-day combination, the user never filled out a PDD. Figure 1 shows MD4, the number of user-day combinations in the dataset by days since app installation. The value 0 indicates that the app was installed the same day. The number of user-day combinations in the dataset decrease rapidly, indicating that many users did not show up in the dataset after only a short time of app use, likely having uninstalled the app.

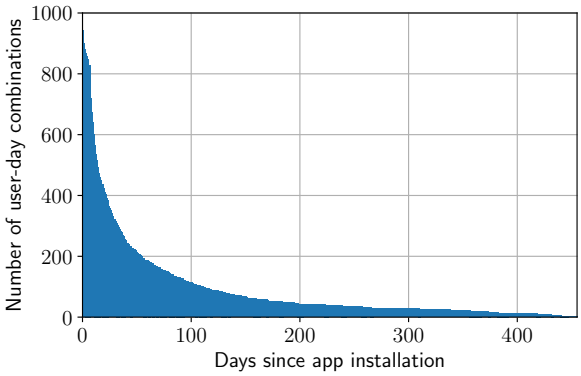

Figure 1: Days since app installation feature showing how many user-day combination we recorded where the user was using the app for the number of days shown on the x-axis.

Note on sampling rate. When collecting sensor data, one can apply different sampling rates. For example, we could record hourly values. For using any specific category of apps, we would then have 24 values for each category (example: one column would be *messaging app usage between 2pm and 3pm*). The more fine-grained the sampling rate, the more precise it captures the behavior of the user. At the same time, the more fine-grained the sampling rate, the more sparse will the resulting data matrix be. We ran the experiments described in the following with different sampling rates (hourly, 6-hour bins, daily values) to see if the prediction performance changes. The performance differences were very small, so we will describe the results for the daily values.

Note on limiting the entries per user. There are some outliers with respect to the overall number of PDDs that a user filled out which could lead to a so-called *compliance bias* (see, e.g., van Berkel, Goncalves, Hosio, Sarsenbayeva, Velloso and Kostakos, 2020). Only four users filled out more than 50 PDDs. Using all data for prediction, each of these four users would account for much more data than several of the users who only submitted few PDDs. Unnikrishnan, Shah, Schleicher, Strandzheva, Dimitrov, Velikova, Pryss, Schobel, Schlee and Spiliopoulou (2020) conducted a study to investigate to what extent machine learning models trained on users with many entries in an EMA system can be used to predict future properties of users with few entries. Their results indicate that separate models perform better, however, their results are based on a rather small sample of 11 users. We conducted our own experiments on our data and, in our case, found only negligible differences between when using a subset of similar users compared to using all data. Thus, we think the compliance bias does not affect the prediction in our case and we use all available data in the following.

Table 2
Datasets overview.

Dataset	# Users	# Rows	# Features	% positive class
DS1–SD+MD	942	42,827	115	5.2%
DS1–MD	942	42,827	4	5.2%
DS1–SD	942	42,827	111	5.2%
DS2	794	794	40	30.9% (at least one PDD, RQ2)

DS1–SD+MD: Includes all sensor data features and all metadata features.
DS1–MD: Exclusively the metadata features.
DS1–SD: Exclusively the sensor data features.

3.2.2. Prediction of filling out at least one PDD

For RQ2, we want to see if we can predict if an onboarding user is going to fill out at least one PDD. We make these predictions based on initial information about the user, as opposed to mobile sensing data. The data we used here, are: (a) demographic information, (b) Big Five personality traits, and (c) phone properties (price (MSRP) and year of release). (a) and (b) were assessed via questionnaires. (a) contains age, sex, use of second phone (y/n), and highest completed level of education. For both (a) and (b), we additionally added metadata: start time, end time, and duration of filling out the questionnaire, binary value if the questionnaire was filled out on a weekday or weekend. For (c), we recorded the phones model name, and looked up MSRP and release year from GSM Arena[3]. The price information might not be the price the user actually paid, nevertheless, it indicates a general idea if the model is a high-end, mid, or low-end model. For 9% of the phones, we did not have the MSRP or release year (missing values).

3.3. Final Datasets

Table 2 shows the final datasets. DS1 contains sensor data (SD) and/or metadata (MD) features for the daily prediction of filling out PDD. DS2 contains features about those users who filled out the demographic and the Big Five questionnaires. The mean age was 34.1 years (SD: 12.8). Regarding sex distribution, the sample is mostly male, with 181/794 females (22.8%). Of those that filled out at least one PDD, the mean number of filled-out PDDs is 6.28 (SD: 10.77). Most users did not fill out many daily questionnaires. Of those using the PDD feature, 47.3% (116/245) filled out only one, two, or three PDDs.

3.4. Machine Learning approaches and evaluation metric(s)

For the given tabular data, tree ensemble methods have been shown to outperform deep learning approaches (Grinsztajn, Oyallon and Varoquaux, 2022; Shwartz-Ziv and Armon, 2022). We tested a variety of tree ensemble methods, and LGBM (Light Gradient Boosting Machine) (Ke, Meng, Finley, Wang, Chen, Ma, Ye and Liu, 2017) performed best. We used the scikit-learn pipeline[4] construct for running our experiments. The pipeline consists of three steps: imputation, scaling, and LGBM. The pipeline construct furthermore makes sure that splitting in train/test set is applied before any of the steps are executed. We performed the hyperparameter search via the scikit-optimize BayesSearchCV[5]. As part of the hyperparameter search, we not only tried different hyperparameters of LGBM, but also searched for the optimal scaling strategy; see Table 3 for an overview of the hyperparameter search space. In order to see if our results are robust, i.e., independent of the randomness of the train/test split, we conducted a 5-fold nested cross-validation. If the overall approach is robust, we expect each model/fold to perform similarly. The cross-validations are set up by StratifiedGroupKFold[6], with the entries of one user representing one group. This avoids a potential bias that would be created when using the same user in both train and test set. StratifiedGroupKFold ensures that not only each group is either in train or test set, but also ensures stratification, i.e., that the percentage of positive cases is the same in training and test set. We use stratification, because otherwise, without stratification, because of the low number of positive cases (see Table 2), there is the chance that the test set might not even have any positive case, which, in turn, could lead the classifier to be trained to always predict the negative class, i.e., not filling out PDD.

The models take the input features and yield probabilities for binary classification (fill out PDD or not fill out PDD). The *threshold* defines which probability is necessary to be classified as *fill out*. To judge how accurate a model is, two metrics are common: precision and recall. Precision describes how many of the items classified as *fill out* are actual *fill out* events (i.e., how many of the selected cases are relevant). Recall describes how many of all *fill out*s were classified as *fill out*s (i.e., how many of the relevant cases are selected). Both precision and recall are depended on the threshold. By default, the threshold is set to 0.5, meaning that the model classifies a sample as *fill out* if the final prediction value is 0.5 or higher. Thus, the threshold can be seen as another hyperparameter to optimize the model's performance. One could imagine optimizing for precision (focusing on being correct about those that are classified as *fill out*, i.e., focusing on avoiding misclassifications) or for recall (i.e., focusing on recognizing all existing *fill out*

[3] https://www.gsmarena.com/, data crawled 2019-10-24.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html, accessed 2023-05-23.

[5] https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html, accessed 2023-05-23.

[6] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedGroupKFold.html, accessed 2023-05-23.

Table 3
Hyperparameter search space for the machine learning pipelines.

Imputer	
	SimpleImputer
Scaler	
	StandardScaler(with_mean=True, with_std=True)
	StandardScaler(with_mean=False, with_std=False)
	MinMaxScaler
	MaxAbsScaler
LGBM	
	deterministic: True
	force_row_wise: True
	learning_rate: [0.01, 0.1, 0.2, 0.3]
	max_depth: [-1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
	n_estimators: [25, 50, 100, 200, 300, 400, 500, 1000, 1500]
	num_leaves: [2, 4, 6, 8, 16, 32, 64, 128, 256]

events). The precision-recall-curve allows to visualize all possible thresholds at the same time and is often used for assessing the quality of a model build on imbalanced data (Saito and Rehmsmeier, 2015; Ozenne, Subtil and Maucort-Boulch, 2015). We are using the precision-recall-curve as our main evaluation metric and report the area under the curve (AUC) in order to have an easy single-value metric.

4. Results

For RQ1, we use daily data to predict if the TYDR user filled out PDD in the evening. In Figure 2, we present the precision-recall curve for the evaluation based on dataset DS1–SD+MD containing both sensor data and metadata features. The baseline performance is equal to the percentage of positive examples, i.e., 5%. With AUC values between 0.85 to 0.89, the performance is very good for all five folds of the nested cross-validation, indicating that the model is robust.

In Figure 3, we present the results for the evaluation with dataset DS1–MD, containing only the four metadata features (cf. Section 3.2.1). The AUC values for the five folds range from 0.88 to 0.91, indicating a slightly better performance overall compared to DS1–SD+MD.

Figure 4 shows the results using dataset DS1–SD, which only contains the sensor data. The AUC values range from 0.14 to 0.21, showing a larger difference between the folds compared to the other two datasets. This indicates that the model is not as robust, i.e., depending on the randomly chosen training and testing data, the performance differed more. While each of the five models still outperforms the baseline of 0.05, the results show only low predictive performance.

For RQ2, we try to predict if a TYDR user is going to fill out at least one PDD, based on demographic data, personality trait data, and smartphone properties. Figure 5 shows the precision-recall-curves for the five models of the nested cross-validation. The performance (AUC) ranges from 0.46 to 0.57, while the baseline is 0.31. We observe a larger variance between the folds, indicating that the performances depended somewhat on the training/test split. While

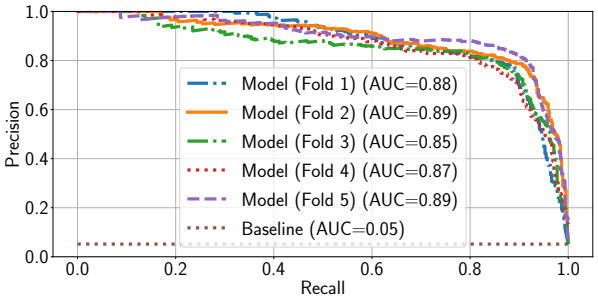

Figure 2: Precision-recall curve on DS1-SD+MD (sensor data and metadata features) for predicting filling out the daily questionnaire in the evening.

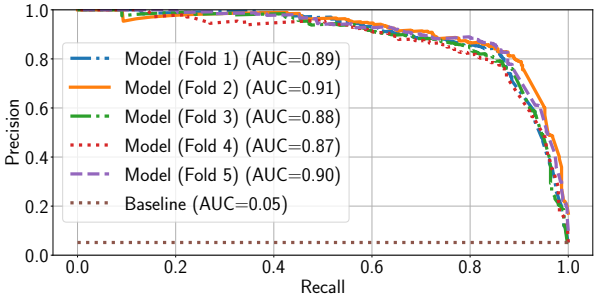

Figure 3: Precision-recall curve on DS1-MD (metadata features only) for predicting filling out the daily questionnaire in the evening.

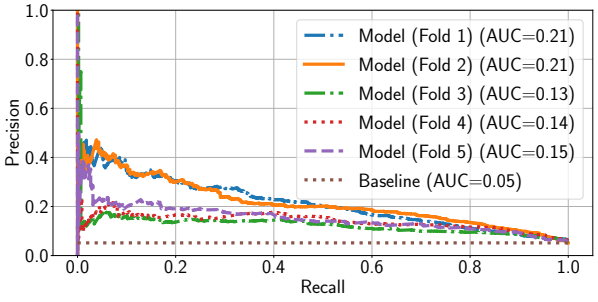

Figure 4: Precision-recall curve on DS1-SD (sensor data features only) for predicting filling out the daily questionnaire in the evening.

the models outperform the baseline, the prediction is not very good.

5. Discussion

First of all, we note that the data collection itself is quite hard. Depending on the category of data, we observe quite a few missing values. The main reason is likely the fragmentation of the Android ecosystem, with different OS versions, OS adaptations by the manufacturer, and different sensors making it difficult to consistently collect high quality sensor data. However, we were still able to capture a complete

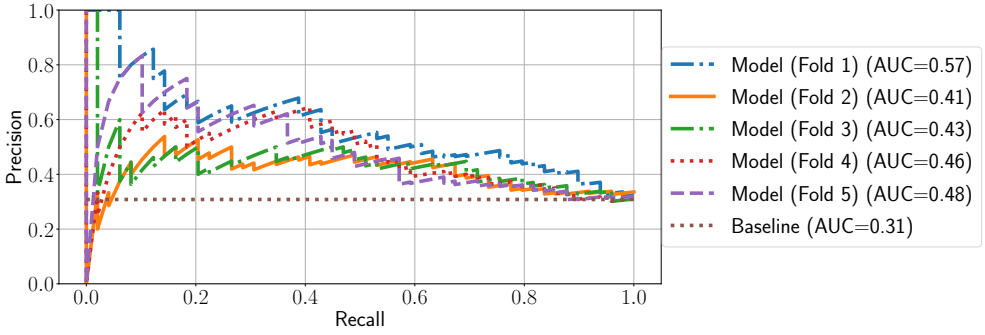

Figure 5: Precision-recall curve on DS2 for predicting filling out at least one daily questionnaire.

picture of the user and his/her activities by having features relating to WiFi, battery, display state, call metadata, photo metadata, step count and duration, and app usage. While a lot of the related work was done on different user groups with specific health conditions (see Section 2), we focused on the general population. Our study is based on comparatively large datasets, and while we expect our results to generalize for other studies conducted on the general population, further studies are needed to confirm our findings.

Regarding RQ1, predicting if a user will fill out a questionnaire in the evening, the predictions based on metadata alone (days since first/last questionnaire filled out, number of questionnaires filled out overall, days passed since app installation) are very good. Intuitively, this shows us that past behavior predicts future behavior. The predictions on sensor data alone are better than chance but nowhere near an accurate prediction. The combination of sensor data and metadata performs slightly worse than metadata features alone. We suspect that this the large amount of sensor data features, 111, introduces noise that makes the model perform worse. Conducting more experiments with more elaborate machine learning pipelines or different models might yield a slight improvement of performance when using sensor data, but likely, the trade-off between collecting data and improvement of performance is not worth it. Collecting and processing noisy sensor data does not improve the performance of the daily prediction, so the effort is not worth it for this purpose. Our finding that past behavior helps better predict future behavior in terms of adherence to EMA prompts is in line with previous research (Schleicher et al., 2020; Jones et al., 2021a; Klaus et al., 2022).

Predicting if a new user will fill out at least one daily questionnaire (RQ2) could be helpful for clinicians and psychologists when recruiting participants or for establishing measures specifically for those we predict to not fill out any questionnaire. Our experiments show, however, that the prediction is not easily possible. The larger range of the values of the area under the precision-recall-curve indicates that more data might be needed to draw more robust conclusions. In contrast to some of the related work, which reported no prediction based on person-level characteristics (Schleicher et al., 2020; Williams-Kerver et al., 2021), our work shows that the prediction is likely to be better than chance. We might have different results compared to the related work because we likely recorded more data points at registration, including, for example, personality traits.

There are some limitations to our study. There might be some bias with respect to the user base. For other types of ecological momentary assessments apps, there might be different relationships between sensor data, metadata, and fill-out behavior. We think that it is unlikely that iPhone users show a different behavior pattern, given that the personality of smartphone users only differs slightly between Android and iOS users (Götz, Stieger and Reips, 2017). Nevertheless, more studies are needed to confirm this and show if iPhone users show a different fill-out behavior. With sensor data from iPhones, the predictive quality of a machine learning model might be higher, given that there are fewer phone models and hard- and software differences between iPhones compared to Android phones. Lastly, there might be other factors that can well predict fill-out behavior that we did not capture with the sensor data that we collected.

Overall, we conducted a study based on a large dataset and showed that sensor data was not good at predicting fill-out behavior of daily questionnaires. Past user behavior on the other hand could predict future fill-out behavior well. More studies are needed regarding the prediction if a new user will fill out at least one daily questionnaire.

CRediT authorship contribution statement

Felix Beierle: Conceptualization, Methodology, Formal Analysis, Funding Acquisition, Writing - Original Draft, Writing - Review & Editing. **Wepan Chada:** Methodology, Formal Analysis, Writing - Review & Editing. **Akiko Aizawa:** Writing - Review & Editing. **Rüdiger Pryss:** Writing - Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share the data.

Acknowledgments

This work was supported by a fellowship within the IFI program of the German Academic Exchange Service (DAAD).

References

Beierle, F., 2021. Integrating Psychoinformatics with Ubiquitous Social Networking: Advanced Mobile-Sensing Concepts and Applications. Springer International Publishing. doi:10.1007/978-3-030-68840-0.

Beierle, F., Matz, S.C., Allemand, M., 2023. Mobile Sensing in Personality Science, in: Mehl, M.R., Eid, M., Wrzus, C., Harari, G.M., Ebner-Priemer, U.W. (Eds.), Mobile Sensing in Psychology: Methods and Applications. Guilford Press, New York, NY, USA. chapter 20, pp. 479–502.

Beierle, F., Tran, V.T., Allemand, M., Neff, P., Schlee, W., Probst, T., Pryss, R., Zimmermann, J., 2018a. Context Data Categories and Privacy Model for Mobile Data Collection Apps. Procedia Computer Science 134, 18–25. doi:10.1016/j.procs.2018.07.139.

Beierle, F., Tran, V.T., Allemand, M., Neff, P., Schlee, W., Probst, T., Pryss, R., Zimmermann, J., 2018b. TYDR – Track Your Daily Routine. Android App for Tracking Smartphone Sensor and Usage Data, in: 2018 IEEE/ACM 5th International Conference on Mobile Software Engineering and Systems (MOBILESoft), ACM. pp. 72–75. doi:10.1145/3197231.3197235.

Beierle, F., Tran, V.T., Allemand, M., Neff, P., Schlee, W., Probst, T., Zimmermann, J., Pryss, R., 2020. What data are smartphone users willing to share with researchers? Journal of Ambient Intelligence and Humanized Computing 11, 2277–2289. doi:10.1007/s12652-019-01355-6.

Boubiche, D.E., Imran, M., Maqsood, A., Shoaib, M., 2019. Mobile crowd sensing – Taxonomy, applications, challenges, and solutions. Computers in Human Behavior 101, 352–370. doi:10.1016/j.chb.2018.10.028.

Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., Srivastava, M.B., 2006. Participatory Sensing, in: Workshop on World-Sensor-Web (WSW'06): Mobile Device Centric Sensor Networks and Applications, ACM. pp. 117–134.

Colombo, D., Cipresso, P., Alvarez, J.F., Palacios, A.G., Riva, G., Botella, C., 2018. An Overview of Factors Associated with Adherence and Dropout to Ecological Momentary Assessments in Depression. Annual Review of Cybertherapy and Telemedicine 16, 11–17.

Fleeson, W., 2001. Toward a Structure-and Process-Integrated View of Personality: Traits as Density Distributions of States. Journal of Personality and Social Psychology 80, 1011–1027. doi:10.1037/0022-3514.80.6.1011.

Gershon, A., Kaufmann, C.N., Torous, J., Depp, C., Ketter, T.A., 2019. Electronic Ecological Momentary Assessment (EMA) in youth with bipolar disorder: Demographic and clinical predictors of electronic EMA adherence. Journal of Psychiatric Research 116, 14–18. doi:10.1016/j.jpsychires.2019.05.026.

Götz, F.M., Stieger, S., Reips, U.D., 2017. Users of the main smartphone operating systems (iOS, Android) differ only little in personality. PLOS ONE 12, e0176921. doi:10.1371/journal.pone.0176921.

Grinsztajn, L., Oyallon, E., Varoquaux, G., 2022. Why do tree-based models still outperform deep learning on typical tabular data?, in: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 507–520.

Jones, S.E., Moore, R.C., Pinkham, A.E., Depp, C.A., Granholm, E., Harvey, P.D., 2021a. A cross-diagnostic study of adherence to ecological momentary assessment: Comparisons across study length and daily survey frequency find that early adherence is a potent predictor of study-long adherence. Personalized Medicine in Psychiatry 29–30, 100085. doi:10.1016/j.pmip.2021.100085.

Jones, S.L., Hue, W., Kelly, R.M., Barnett, R., Henderson, V., Sengupta, R., 2021b. Determinants of Longitudinal Adherence in Smartphone-Based Self-Tracking for Chronic Health Conditions: Evidence from Axial Spondyloarthritis. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 5, 16:1–16:24. doi:10.1145/3448093.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.

Klaus, F., Peek, E., Quynh, A., Sutherland, A.N., Selvam, D., Moore, R.C., Depp, C.A., Eyler, L.T., 2022. Mobile survey engagement by older adults is high during multiple phases of the COVID-19 pandemic and is predicted by baseline and structural factors. Frontiers in Digital Health 4.

Kraft, R., Schlee, W., Stach, M., Reichert, M., Langguth, B., Baumeister, H., Probst, T., Hannemann, R., Pryss, R., 2020. Combining Mobile Crowdsensing and Ecological Momentary Assessments in the Healthcare Domain. Frontiers in Neuroscience 14. doi:10.3389/fnins.2020.00164.

McCrae, R.R., John, O.P., 1992. An Introduction to the Five-Factor Model and Its Applications. Journal of Personality 60, 175–215.

Murray, A., Yang, Y., Zhu, X., Speyer, L., Brown, R., Eisner, M., Ribeaud, D., 2023a. Respondent characteristics associated with adherence in a general population ecological momentary assessment study. International Journal of Methods in Psychiatric Research n/a, e1972. doi:10.1002/mpr.1972.

Murray, A.L., Brown, R., Zhu, X., Speyer, L.G., Yang, Y., Xiao, Z., Ribeaud, D., Eisner, M., 2023b. Prompt-level predictors of compliance in an ecological momentary assessment study of young adults' mental health. Journal of Affective Disorders 322, 125–131. doi:10.1016/j.jad.2022.11.014.

Ozenne, B., Subtil, F., Maucort-Boulch, D., 2015. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. Journal of Clinical Epidemiology 68, 855–859. doi:10.1016/j.jclinepi.2015.02.010.

Saito, T., Rehmsmeier, M., 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE 10, e0118432. doi:10.1371/journal.pone.0118432.

Schlee, W., Pryss, R.C., Probst, T., Schobel, J., Bachmeier, A., Reichert, M., Langguth, B., 2016. Measuring the Moment-to-Moment Variability of Tinnitus: The TrackYourTinnitus Smart Phone App. Frontiers in Aging Neuroscience 8. doi:10.3389/fnagi.2016.00294.

Schleicher, M., Unnikrishnan, V., Neff, P., Simoes, J., Probst, T., Pryss, R., Schlee, W., Spiliopoulou, M., 2020. Understanding adherence to the recording of ecological momentary assessments in the example of tinnitus monitoring. Scientific Reports 10, 1–13. doi:10.1038/s41598-020-79527-0.

Shiffman, S., Stone, A.A., Hufford, M.R., 2008. Ecological Momentary Assessment. Annual Review of Clinical Psychology 4, 1–32. doi:10.1146/annurev.clinpsy.3.022806.091415.

Shwartz-Ziv, R., Armon, A., 2022. Tabular data: Deep learning is not all you need. Information Fusion 81, 84–90. doi:10.1016/j.inffus.2021.11.011.

Stone, A.A., Shiffman, S., 1994. Ecological Momentary Assessment (EMA) in Behavioral Medicine. Annals of Behavioral Medicine 16, 199–202. doi:10.1093/abm/16.3.199.

Unnikrishnan, V., Shah, Y., Schleicher, M., Strandzheva, M., Dimitrov, P., Velikova, D., Pryss, R., Schobel, J., Schlee, W., Spiliopoulou, M., 2020. Predicting the Health Condition of mHealth App Users with Large Differences in the Number of Recorded Observations - Where to Learn from?, in: Appice, A., Tsoumakas, G., Manolopoulos, Y., Matwin, S. (Eds.), Discovery Science, Springer International Publishing. pp. 659–673. doi:10.1007/978-3-030-61527-7_43.

van Berkel, N., Goncalves, J., Hosio, S., Sarsenbayeva, Z., Velloso, E., Kostakos, V., 2020. Overcoming compliance bias in self-report studies: A cross-study analysis. International Journal of Human-Computer Studies 134, 1–12. doi:10.1016/j.ijhcs.2019.10.003.

Virginia Anikwe, C., Friday Nweke, H., Chukwu Ikegwu, A., Adolphus Egwuonwu, C., Uchenna Onu, F., Rita Alo, U., Wah Teh, Y., 2022. Mobile and wearable sensors for data-driven health monitoring system: State-of-the-art and future prospect. Expert Systems with Applications 202, 117362. doi:10.1016/j.eswa.2022.117362.

Wen, C.K.F., Schneider, S., Stone, A.A., Spruijt-Metz, D., 2017. Compliance With Mobile Ecological Momentary Assessment Protocols in Children and Adolescents: A Systematic Review and Meta-Analysis. Journal of Medical Internet Research 19, e132. doi:10.2196/jmir.6641.

Williams-Kerver, G.A., Schaefer, L.M., Hazzard, V.M., Cao, L., Engel, S.G., Peterson, C.B., Wonderlich, S.A., Crosby, R.D., 2021. Baseline and momentary predictors of ecological momentary assessment adherence in a sample of adults with binge-eating disorder. Eating Behaviors 41, 101509. doi:10.1016/j.eatbeh.2021.101509.

Zimmermann, J., Woods, W.C., Ritter, S., Happel, M., Masuhr, O., Jaeger, U., Spitzer, C., Wright, A.G.C., 2019. Integrating structure and dynamics in personality assessment: First steps toward the development and validation of a personality dynamics diary. Psychological Assessment 31, 516–531. doi:10.1037/pas0000625.